

Практическая работа №3. Отчёт

Команда №1

Состав: Азимжанова Инаара (группа 312), Банников Дмитрий (группа 311), Ройтман Андрей (группа 311).

Задача состояла в следующем:

1. Считать данные из training.xlsx. Проверить, является ли ряд стационарным в широком смысле. Это было сделано с помощью проведения теста Дики-Фуллера. (25 баллов)
2. Разложить временной ряд на тренд, сезональность, остаток в соответствии с аддитивной, мультипликативной моделями. Визуализировать их, оценить стационарность полученных рядов, сделать выводы. (15 баллов)

Используемые программные средства:

- Python — интерпретируемый язык программирования, используемый для реализации задания. В реализации применялись следующие библиотеки:
 - *numpy (np)* — библиотека, поддерживающая работу с многомерными массивами
 - *pandas (pd)* — библиотека, предоставляющая инструменты для обработки и анализа данных (в реализации используется для работы с временными рядами)
 - *matplotlib.pyplot (plt)* — библиотека, предоставляющая инструменты для построения графиков. В частности, использовалась операция изменения параметров *rcParams* из процедурного интерфейса *pylab* библиотеки *matplotlib*.
 - *statsmodels.api (sm)* — библиотека, предоставляющая инструменты для статистических вычислений
- Jupyter Notebook — основанный на web программный интерфейс, используемый в работе для написания кода на Python

Теоретическая постановка задачи

Определение. Временной ряд — это совокупность наблюдений экономической величины в различные моменты времени.

Временной ряд обладает двумя параметрами: периодом времени и значениями показателя (уровни ряда).

Определение. Временной ряд y называется *стационарным*, если $E[y] = Const$, $D[y] = Const$, $Cov(y_t, y[t-k]) = const * k$, то есть если эти параметры не зависят от времени. Другое определение стационарности — это отсутствие тренда.

Определение. *Тренд* временного ряда — это изменение, определяющее общее направление развития.

То есть это общее направление графика ряда (возрастает/убывает/не изменяется)

Для *визуальной оценки* данный график проверяется на наличие тренда. Тренд есть — ряд нестационарный, тренда нет — ряд стационарен. Стоит отметить, что эта оценка довольно грубая и во многих случаях из внешнего вида графика нельзя сказать, стационарен ряд или нет.

Определение. *Скользящая средняя (Moving Average - MA)* — среднее арифметическое значений исходной функции за установленный период. Скользящая средняя сглаживает краткосрочные колебания и помогает выделить основные тенденции.

Определение. *Стандартное отклонение* — показывает, на сколько в среднем отклонился ряд от средней вариации ряда (от среднего арифметического, в нашем случае).

Скользящая средняя вместе со стандартным отклонением составляют *скользящие статистики*.

Тест Дики-Фуллера

Тест Дики-Фуллера используется для проверки ряда на стационарность. Он проверяет ряд на наличие единичных корней.

Определение. Временной ряд имеет хотя бы один *единичный корень*, если его первые разности образуют стационарный ряд.

Обозначение. $y(t) \sim I(1)$, или $\Delta y(t) = y(t) - y(t-1) \sim I(0)$, где Δ — разностный оператор, $I(j)$ — означает, что ряд является интегрированным порядка j , $I(0)$ — ряд стационарен.

Тест Дики-Фуллера проверяет значение коэффициента a в авторегрессионном уравнении 1-го порядка. Оно имеет вид:

$y(t) = a * y(t-1) + \varepsilon(t)$, $\varepsilon(t)$ — ошибка.

- $a = 1$ — есть единичные корни, стационарности нет
- $|a| < 1$ — нет единичных корней, есть стационарность
- $|a| > 1$ — не свойственно для временных рядов, которые встречаются в реальной жизни — требуется более сложный анализ.

Преобразуем уравнение:

$y(t) = a * y(t-1) + \varepsilon(t) \Rightarrow y(t) - y(t-1) = a * y(t-1) - y(t-1) + \varepsilon(t) \Rightarrow \Delta y(t) = (a-1) * y(t-1) + \varepsilon(t) \Rightarrow \Delta y(t) = b * y(t-1) + \varepsilon(t)$, где $b = a-1$.

- Основная гипотеза: $H_0: b = 0$ — процесс не стационарен.
- Альтернативная гипотеза: $H_1: b < 0$ — процесс стационарен.

Определение. Достоверность статистики — мера уверенности в "истинности" результата.

Чем достоверность меньше, тем больше доверия. Как только она становится больше определённых критических значений, становится ясно, что единичные корни существуют и ряд не стационарен.

Определение. *Уровень значимости* есть степень отклонения от гипотезы, обычно в процентном выражении.

То есть, если наша достоверность превысила 5% уровень значимости, то гипотеза отвергается, то есть процесс не стационарен.

Тренд, сезонность, остаток. Аддитивная и мультипликативная модели.

Определение. *Трендом* временного ряда называется изменение, определяющее общее направление развития.

То есть это общее направление графика ряда (возрастает/убывает/не изменяется).

Определение. *Сезональностью* временного ряда называются периодические колебания, наблюдаемые во временных рядах.

Определение. *Остатком* временного ряда называется разница между предсказанным и наблюдаемым значением.

- Общий вид *аддитивной модели*: $Y = T + S + E$;
- Общий вид *мультипликативной модели*: $Y = T * S * E$;
- T — тренд, S — сезонность, E — остаток

Аддитивная модель.

Для вычисления сезонности необходимо найти скользящее среднее, от скользящего среднего найти еще раз скользящее среднее - получим центрированное скользящее среднее.

Сезональность — это разность временного ряда и центрированного скользящего среднего.

Тренд на позволяет найти метод наименьших квадратов, который приближает временной ряд.

Остаток — временной ряд, из которого вычитается тренд и сезонность.

Мультипликативная модель.

Аналогично с аддитивной моделью находим центрированное скользящее среднее.

Сезональность есть частное временного ряда и центрированного скользящего среднего. Тренд находится аналогично.

Остаток — частное временного ряда и частного сезонности и тренда.

Исполнение

Описание хода программы

Сперва мы извлекаем с помощью функции `pd.read_excel('training.xlsx', index_col='Date')` из файла `training.xlsx` в виде пар данных ячеек столбца `date` и `value` создаётся объект `train` типа `DataFrame`. Затем выполняется первый пункт задания, то есть с помощью теста Дики-Фуллера на устойчивость в широком смысле проверяется временной ряд, записанный в файле. Тестирование проходит в написанной функции `diki(train)`.

Следующим действием идёт подготовка к выполнению второго пункта задания и завершение первого, то есть с помощью функции `figure` библиотеки `plt` создаётся пространство размером 20 на 10 дюймов. После мы выделяем на всё пространство один график, содержащий данные таблицы `train`, и рисуем его с помощью `plt.show()`.

Затем выполняется непосредственно второй пункт задания. Из библиотеки `pylab` импортируется упомянутая в описании задачи функция `rcParams` для изменения размера пространства для графиков с 20 на 10 до 11 на 9 дюймов. Функцией `sm.tsa.seasonal_decompose(train, model='additive')` вычисляется и возвращается класс, содержащий в качестве собственных объектов сезональность, тренд и остаток в аддитивной модели в применении к временному ряду в `train`. Результат сохраняется в переменной `decomposition`. Следующим действием мы добавляем полученные ряды в пространство графиков `plt`, и проверяем каждый ряд с помощью теста Дики-Фуллера на стационарность функцией `diki()`. Затем аналогичные действия выполняются после вызова `sm.tsa.seasonal_decompose(train, model='multiplicative')`, только перед этим пространство для графиков создаётся вновь для отдельного отображения результатов работы мультипликативной модели.

`diki(train)`

Используемые библиотеки. `statsmodels.api`, `pandas` (функции `diff` и `dropna`).

Ввод. На ввод подаётся объект типа `DataFrame`.

Задача. Проверить вводимый ряд на стационарность.

Возвращаемое значение. Функция возвращает 1, если ряд не стационарен, и 0 в противном случае

Работа программы. Сперва функцией `pandas.Series.diff(periods = 1)` мы вычисляем разности между соседними значениями наблюдений в объекте `train`, а элементы, что не имеют в результате значений, удаляются функцией `pandas.dropna()`. Заметим, что для первой функции объект `train` сперва приводится к типу `Series`, а

затем для второй функции обратно к типу *DataFrame*. Затем полученную последовательность мы проверяем на стационарность функцией *statsmodels.tsa.stattools.adfuller()*, которая возвращает массив с различными показателями, из которых мы сравниваем первый, то есть достоверность статистики (или *p-value*), и второй внутренний элемент четвёртого, то есть критические значения с уровнем значимости 5%. Если больше достоверность, то ряд не стационарен, в противном случае — нет, и выводятся соответствующие сообщения и возвращаются уже указанные значения.